ЗК

УДК 681.3.01

И.П. Кузнецов, М.М. Шарнин, А.Г. Мацкевич
Институт проблем информатики РАН, г. Москва, Россия
Россия, 117900, г. Москва, ул. Вавилова 44, кор. 2, igor-kuz@mtu-net.ru

Технология извлечения структур знаний с использованием аппарата расширенных семантических сетей

I.P. Kuznetsov, M.M. Sharnin, A.G. Matskevich
Institute of Informatics Problems of RAS, c. Moscow, Russia
Russia, 117900, Moscow, Vavlova str. 44, building 2, igor-kuz@mtu-net.ru

Technology of Knowledge Extraction on the Base of Extended Semantic Networks

І.П. Кузнєцов, М.М. Шарнін, А.Г. Мацкевич
Інститут проблем інформатики РАН, м. Москва, Росія
Росія, 117900, м. Москва, вул. Вавілова 44, кор. 2, igor-kuz@mtu-net.ru

Технологія відтворення структур знань з використанням апарату розширених семантичних мереж

В статье рассматривается задача извлечения из текстов естественного языка структур знаний: информационных объектов («именованных сущностей»), их свойств, связей и фактов участия в действиях. Для этих целей разработан инструментарий: язык представления знаний (расширенные семантические сети – РСС) и их обработки (язык преобразования структур – ДЕКЛ). На этой основе созданы технологии, которые обладают следующими особенностями. Из текстов извлекаются не отдельные объекты (именованные сущности), а структуры знаний, представляющие связи объектов и их участие в действиях и событиях. Для извлечения структур знаний разработан уникальный семантико-ориентированный лингвистический процессор (ЛП), осуществляющий глубинный анализ текстов ЕЯ и выявляющий десятки типов объектов вместе с их структурами. Процессор ЛП управляется лингвистическими знаниями, представляющими собой декларативные структуры и обеспечивающие быструю настройку ЛП на предметную область и язык. Основой лингвистических знаний являются правила, обладающие высокой степенью избирательности при выявлении объектов («сущностей»), средствами устранения коллизий при их применении. Это позволяет минимизировать шумы и потери.
Ключевые слова: извлечение знаний из текстов, лингвистические процессоры, расширенные семантические сети, обработка структур знаний.

The paper is devoted to the extracting of knowledge structures from the natural language texts, i.e. information objects (“Named Entities”), their features, relationships, and participation in the actions and events. For this purpose, the language used for knowledge representation (extended semantic networks/ESN) and tools for processing (language for structure conversion LSC) are considered. On this base, the new technologies are proposed. These technologies have the following features: extraction from the texts of knowledge structures that represent the links of named entities and their participation in actions and events. For the knowledge extraction the unique semantic-oriented language processor (LP) are designed. Processor LP provides the deep analysis of NL-texts and revealing set of objects together with their structures. Processor LP is controlled by the linguistic knowledge, which are declarative structures (on ESN) and which provides the quick tuning of LP on subject area and language, both Russian and English.
Key Words: knowledge extraction from texts, semantic-oriented linguistic processor, extended semantic networks, knowledge structure processing.

У статті розглядається задача знайдення у текстах природної мови структур знань: інформаційних об'єктів («іменованих сутностей»), їх якостей зв'язків і фактів участі у діях. Для цих цілей розроблений інструментарій: мова представлення знань (розширені семантичні мережі – РСМ) та їх обробки (мова перетворення структур – ДЕКЛ). На цій основі створені технології, що мають наступні особливості.
З тестів виділяються не окремі об'єкти (іменовані сутності), а структури знань, що представляють зв'язки об'єктів та їх участь у діях та подіях. З метою виділення структур знань розроблений винятковий семантико-орієнтований лінгвістичний процесор (ЛП), що здійснює глибинний аналіз текстів ЕЯ та виявляє десятки типів об'єктів разом з їх структурами. Процесор ЛП керується лінгвістичними знаннями, які представляють собою декларативні структури та забезпечують швидке настроювання ЛП на предметну сферу та мову. Основою лінгвістичних знань є правила, що мають високий ступінь вибірковості при виявленні об'єктів («сутностей»), засобами усунення колізій при їхньому використанні. Це дозволяє мінімізувати шуми та втрати.
Ключові слова: знайдення знань з текстів, лінгвістичні процесори,
поширені семантичні мережі, обробка структур знань.

Введение

В настоящее время проблема извлечения знаний становится все более актуальной, что связано с развитием сети Интернет, где накапливаются громадные объемы информации. В основном, это тексты на естественном языке (ЕЯ). Для избирательного извлечения информации по запросам пользователя требуется привлекать семантические отношения и компоненты. В связи с этим все большее распространение и развитие получают такие направления, как семантический WEB, языки RDF (для представления отношений), OWL (для представления онтологий) и др. [www.semantictools.ru].

Одно из направлений связано с извлечением из текстов ЕЯ, так называемых, информационных объектов (лиц, организаций, адресов, дат и др.) и связей между ними. Другое название объектов – «named entities» (NE) или «именованные сущности» [1], [2]. Наиболее продвинутые системы извлечения сущностей разработаны в Станфордском университете (Stanford NER system), Иллинойском университете (Illinois NER system), а также «Lingpipe NER system» и др. Такие системы, как правило, ориентированы на выделение нескольких типов именованных сущностей. Например, первая система типа 7 class выделяет только 7 типов сущностей. Более того, во многих системах не учитываются связи. Их работа заканчивается лишь разметкой текстов с выделением компонент, соответствующих сущностям (NE). При использовании таких разметок (например, для семантических поисков или аналитических решений) возникают существенные трудности. Среди реально работающих отечественных систем следует отметить «PullEnti», «Semantix» (Синергетические системы), KEYWEN и др., [http://ipiranlogos. com/ru/ Systems/].

В связи со сказанным перспективным представляется направление, когда извлекаются не только объекты, но и их связи, в том числе факты их участия в действиях или событиях. Возникают структуры знаний, обеспечивающие другой уровень решения задач. Но при этом требуются специальные средства представления и обработки знаний.

Для представления структур знаний в рамках проектов ИПИ РАН разработан новый математический аппарат и соответствующий инструментарий: язык расширенных семантических сетей (РСС), а для обработки – продукционный язык ДЕКЛ [3]. Они образует законченный технологический комплекс, ориентированный на сложные задачи, связанные с логическим выводом, преобразованием представлений, лингвистическим анализом, экспертными и аналитическими решениями. На этой основе построено:

– семейство оболочек для построения экспертных (ШЕДЛ, DECSAY и др.);

– множество самих экспертных систем («Токсиколог» – для института Склифосовского, «Тибет» – для лечения методами тибетской медицины и др.);

– несколько лингвистических оболочек для создания языков и организации естественно-языкового общения (ДИЕС, ИКС);

– семейство лингвистических процессоров глубинного анализа текстов для конкретных логико-аналитических систем;

– ряд интеллектуальных систем различного назначения, например, СПРУТ (для выявления организованных преступных формирований), «Криминал» (решает задачи оперативно-аналитической обработки и семантического поиска), «Резюме» (для формализация заявок на работу) и др. [4].

Как показал опыт, разработанный инструментарий позволяет быстро строить интеллектуальные системы высокой степени сложности. В данной статье рассматривается использование этого аппарата для задач извлечения и обработки структур знаний. Успешность систем зависит от извлекаемой информации (количества и типов извлекаемых объектов и связей), а также от способа представления результатов (знаний) и средств их обработки, что непосредственно определяет класс и качество решаемых задач. Имеются в виду задачи идентификации объектов, выявления и анализа фактографической информации, семантического поиска, экспертных решений, ответа на запросы, выраженные на ЕЯ, и др. [4], [5].

Для извлечения знаний требуется разработка соответствующих лингвистических процессоров, отображающих тексты ЕЯ на структуры знаний. При этом формализмы представления знаний должны учитывать высокую степень разнообразия объектов и их связей. Например, для лиц должны быть представлены не только родственные связи и их анкетные данные, но и действия или события, в которых эти лица участвуют. Собственно, они и составляют факты. Такие действия привязаны ко времени, месту. Более того, одни события могут быть составной частью других. Они могут быть связаны причинно-следственными и временными отношениями. Для ряда задач подобные связи играют важную роль. Их тоже нужно выявлять и использовать. Поэтому следует считать, что действия и соответствующие им факты – это тоже информационные объекты, связанные между собой и с другими информационными объектами. Возникают сложные структуры знаний. Для их представления и разработан язык РСС [3].

Для извлечения знаний разработан и постоянно совершенствуется семантико-ориентированный лингвистический процессор (ЛП), анализирующий тексты ЕЯ и автоматически формирующий на этой основе структуры знаний – так называемые содержательные портреты документов (СП-документов) [4], [6], [7]. Они представляются в виде РСС и образуют базу знаний (БЗ), в рамках которой обеспечивается анализ высокой степени глубины и сложности.

Отметим, что первые такие процессоры были разработаны для системы «Криминал», ориентированной на информационную поддержку оперативно-аналитической работы в ГУВД г. Москвы. Система проводит глубинный анализ документов, циркулирующих в ГУВД, выделяет до 40 типов объектов, их свойств, отношений и участие в действиях. В результате автоматически формируется база знаний, которая служит основой для семантических поисков и экспертных решений. Система «Криминал» отлаживалась на 500 тыс. происшествий из сводок ГУВД г. Москвы. По основным объектам удалось добиться хороших результатов: коэффициент шумов в компонентах (лишних слов в объектах) – не более 1 – 2% и потерь (отсутствие нужных слов) – не более 1% [3], [4]. Развитие этих процессоров нашло свое воплощение в системах «Аналитик», «АнтиТеррор» (ИПИ РАН), «Semantix» (совместно с компанией «Синергетические системы») [7], [8]. Рассмотрим технологическую базу этих процессоров [http://ipiranlogos.com/ru/Technologies/].

1 Компоненты семантико-ориентированного лингвистического процессора

Семантико-ориентированный ЛП состоит из четырех основных компонент.

Блок лексико-морфологического анализа (реализован на С++). Выделяет из документа слова и предложения и выдает в виде семантической сети, представляющей пространственную структуру документа (ПС-документа). Эта структура имеет вид линейной последовательности связанных фрагментов, представляющих слова в нормальной форме, числа, знаки, а также их основные признаки – лексические, морфологические и семантические. Для придания словам и словосочетаниям дополнительных семантических признаков используется набор предметных словарей: словарь стран, регионов России, имен, профессий и др. [5], [9].

Блок синтактико-семантического анализа проводит анализ ПС-документа, выделяет объекты и связи. Для этого используются специальные правила анализа структур (п. 4). В результате строится другая семантическая сеть, называемая содержательным портретом документа (СП-документа) [4-6], [10]. Такие портреты образуют структуры знаний, которые запоминаются в базе знаний (БЗ). Блок обеспечивает:

– извлечение информационных объектов (лиц, организаций, событий, их места, …);

– выявление связей объектов. Например, как лица связаны с организациями, адресами и др.;

– анализ глагольных форм, причастных и деепричастных оборотов с выявлением фактов участия объектов в тех или иных действиях;

– идентификацию объектов с учетом анафорических ссылок и сокращенных наименований;

– выявление связей действий с их местом или временем (где и когда имело данное действие или событие).

– анализ причинно-следственных и временных связей между действиями и событиями.

Блок экспертных решений. Анализирует структуры знаний в БЗ, решает логико-аналитические задачи и формирует дополнительную (экспертную) информацию, необходимую для пользователя.

Обратный лингвистический процессор. Преобразует структуры знаний в тексты ЕЯ, которые должны быть выданы пользователю.

Имеется ряд вспомогательных блоков, один из которых – блок построения каталогов объектов. Этот блок выделяет из СП-документов объекты определенного типа, которые упорядочиваются по алфавиту и образуют каталог. Например, таким способом создаются каталоги лиц (их ФИО), дат, адресов и др. – только тех, которые встретились в документах.

Процессор ЛП реализован средствами языка ДЕКЛ и управляется лингвистическими знаниями (ЛЗ) в виде предметных словарей, средств параметрической настройки, а также правил выделения объектов и связей (п. 4). С помощью ЛЗ осуществляется настройка ЛП на соответствующие категории пользователей и корпуса текстов. В результате возникает конкретная реализация. Таким образом, речь идет о средствах построения семейства процессоров ЛП с широкими возможностями их настройки и совершенствования.

С помощью процессоров ЛП из текстов ЕЯ выделяется более 40 типов объектов. Их количество зависит от предметной области и задач пользователя. На рис. 1 представлены типовые объекты, выделяемые ЛП в системах различного назначения.

Отметим, что чем больше таких объектов, тем больше трудностей при их выделении. Дело в том, что правила выделения вступают в коллизии, захватывают чужеродные компоненты вместо своих и т.д. Такие правила должны быть очень дифференцированными, что определяет их конструктивные особенности (п. 4).



Рисунок 1 – Выделяемые информационные объекты

Увидеть графы, составленные из этих объектов и отражающие семантику текстов, можно на сайте [http://ipiranlogos.com/ru/Demo-1/].

2 Содержательные портреты документов

С помощью семантико-ориентированного ЛП из текстов ЕЯ извлекаются информационные объекты и связи, а также конструкции ЕЯ, представляющие связи, действия (факты, события). Они преобразуются в однотипные фрагменты на РСС, имеющие вид:

<тип объекта>(<арг.1>,<арг.2>,…/<код фрагмента>),
<вид связи>(<арг.1>,<арг.2>,…/<код фрагмента>),
<имя действия>(<арг.1>,<арг.2>,…/<код фрагмента>).

Код фрагмента – это константа, которая соответствует объекту или действию, представленному с помощью всего фрагмента. Аргументами (арг. N) могут быть слова в нормальной форме (необходимо для идентификации и поиска), или коды других фрагментов. В результате возникает аппарат (формализм РСС), покрывающий логику предикатов и множество других математических средств. В рамках данного аппарата обеспечивается представление случаев, когда одни объекты включают в себя другие, или когда комплексные действия включают в себя объекты и другие действия. Такие случаи недопустимы в логике предикатов, но являются типичными для текстов ЕЯ, что легко представляется в виде РСС, и соответственно, в БЗ.

Множество таких фрагментов, сформированных на базе текстового документа, составляет структуру знаний – содержательный портрет (СП-документа). Рассмотрим, как выглядят такие структуры в формализме РСС [4], [8], [10].

Пример 2. Текст взят из сводок происшествий ГУВД г. Москвы:

01.02.98 г. в 16-30 в ОВД обратился гр-н Митрофанов Виктор Михайлович, 1955 г.р., прож.: Боровское шоссе 38-211, н/р. Он заявил, что 01.02.98 г. в 10-00 у д. 3 по ул. Федосьино неизвестные, находясь в пьяном виде, учинили скандал, выражались

нецензурной бранью, натравили собаку. В результате чего Митрофанов обратился в травмпункт, где был поставлен диагноз: укус ноги.

Содержательный портрет данного текста (СП-текста) имеет вид:

ДОК_(22,“1-02-98”, “СВОДКА;”/0+) 0-(RUS)
ОВД_(ОВД/1+)
FIO(МИТРОФАНОВ,ВИКТОР,МИХАЙЛОВИЧ,1955/2+)
БЕЗРАБОТНЫЙ(2-/3+) 3-(22,PROP_)
АДР_(БОРОВСКИЙ,Ш.,38,211/4+)
ПРОЖ.(2-,4-)
АДР_(УЛ.,ФЕДОСЬИНО,ДОМ,3/5+)
FIO(" "," "," ",НЕСКОЛЬКО/6+)
НЕИЗВЕСТНЫЙ(6-)
ПЬЯНЫЙ(6-/7+) 7-(2,PROP_)
СКАНДАЛ(6-,ПЬЯНЫЙ/8+) 8-(22,ACT_)
СООБЩИТЬ(2-,8-/9+) 9-(22,ACT_)
ДАТА_(1998,02,~01,"10-00"/10+)
Когда(9-,10-)
ОБРАТИТЬСЯ(1-,2-/11+) 11-(22,ACT_)
ДАТА_(1998,02,~01,"16-30"/12+)
Когда(11-,12-)
ВЫРАЖАТЬСЯ(6-,НЕЦЕНЗУРНЫЙ,БРАНЬ/13+) 13-(22,ACT_)
НАТРАВИТЬ(6-,СОБАКА/14+) 14-(0,ACT_)
ОБРАТИТЬСЯ(2-,В,ТРАВМПУНКТ/14+) 14-(0,ACT_)
ПОСТАВИТЬ(ДИАГНОЗ,УКУС,НОГА/16+) 16-(0,ACT_)
ПРЕДЛ_(22,11-,4-,3-,9-,13-,14-/17+) 17-(2,15,341)
ПРЕДЛ_(22,15-,16-/18+) 18-(6,342,448)

Содержательный портрет состоит из элементарных фрагментов, аргументами которых являются слова в нормальной или канонической форме (например, для существительных – в ед. числе, им. падеже, для прилагательных – дополнительно муж. род и т.д.). Это необходимо для поиска и обработки. Как уже говорилось, каждый элементарный фрагмент имеет свой уникальный код, который записывается в виде числа с знаком «+» и отделяется косой линией. Например, в фрагменте ОВД_(ОВД/1+) знак «1+» есть его код. Знак «1-» – это ссылки на него. Например, в фрагменте ОБРАТИТЬСЯ(1-,2-/11+) знаки «1-» и «2-» означают, что в ОВД обратилось лицо, представленное FIO(МИТРОФАНОВ, … /2+).

Фрагменты типа ДОК_(22,«1-02-98.TXT»,«СВОДКА;»/0+) 0-(RUS) указывают, что содержательный портрет построен на основе русскоязычного текста документа (RUS) с номером 22 из файла 1-02-98.TXT”, который обрабатывался как сводка происшествий (от этого зависят лингвистические знания). Следующие фрагменты представляют: отделение милиции (ОВД_), лицо (ФИО), его свойство (PROP) – безработный, адрес (АДР_) и т.д. Знаки «3+» , «3» , «4+» , «4»… – это коды фрагментов, с помощью которых задаются их связи и отношения. Например, фрагмент ПРОЖ.(2-4) представляет отношение, что лицо (представленное как ФИО с кодом «2+») проживает по адресу (фрагмент АДР_ с кодом «4+»). Действия также представляются в виде фрагментов типа СКАНДАЛ(6-,ПЬЯНЫЙ/8+) 8-(22,ACT_), где представлено, что «лицо (ФИО с кодом “6+”), будучи пьяным, учинило скандал». С помощью кода («8+», «8-») указывается, что фрагмент представляет действие (ACT_) и относится к документу с номером 22. Такие коды также служат для представления времени, места действия и фактов их комбинирования – когда одно действие включено в состав другого. Будем называть такие действия составными. Например, фрагмент

СООБЩИТЬ(2-,8-/9+) представляет, что лицо (код «2+») сообщило о действии (код «8+»), т.е. об «учиненном скандале». Следующие фрагменты ДАТА_(…/10+) Когда (9-,10-) представляют время (ДАТА_) и что оно относится к действию «сообщить» (код «9+»).

Особую роль играют фрагменты ПРЕДЛ_(...), которые соответствуют предложениям. Они заполняются словами, не вошедшими в информационные объекты (в данном примере их нет), а также кодами самих объектов. К этим фрагментам добавляются указатели их местоположения в тексте. Например, фрагмент ПРЕДЛ_(22,11-,3-,9-,13-,14-/17+) 17-(2,15,341) представляет тот факт, что объекты с кодами «11-» (соответствует действию «обратиться»), «3-» (соответствует свойству «безработный») и др. находятся в предложении, которое начинается с 2-ой строки текста документа и занимают место от 15-го байта до 341-го. Это средства позиционирования, которые необходимы для работы обратного ЛП.

Отметим, что вся информация представляется в БЗ на однородной основе, что очень важно для обработки, осуществляемой продукциями языка ДЕКЛ. Левая и правая части таких продукции (правила ЕСЛИ, …ТО) состоят из аналогичных фрагментов, содержащих переменные. Последние обозначаются в процессе применения продукций – сопоставления ее левой части со структурами в БЗ и выполнения действий, указанных в правой части. С помощью продукций осуществляются различные виды преобразования структур знаний, в том числе, осуществляющие разнообразные формы логического вывода, преобразование представлений, экспертные оценки и др. Языки РСС и ДЕКЛ составляют универсальную инструментальную среду, ориентированную на представление и обработку семантической информации, извлекаемой из текстов ЕЯ [http://ipiranlogos.com/ru/Tools/].

3 Принципы выявления объектов и связей

Для выявления многих объектов используются характеристические слова, по которым определяется наличие объекта. Например, слова «дом» (за которым стоит число) или «улица» (за которым стоит слово с большой буквы) определяют наличие объекта типа «адрес». Аналогично, слова «фирма», ООО, «банк» и др. (за которыми стоит слово с большой буквы или слова в кавычках) определяют наличие объекта типа «организация». Это характеристические слова, с которых начинается выделение объекта, включающего эти слова.

При отсутствии характеристических слов используется принцип ожидания – после одних слов или объектов ожидается наличие других. Например, если после слова «инженер» стоит слово с большой буквы (и оно не обладает признаками «организации»), то, скорее всего, оно относится к ФИО. Вместо слова «инженер» может быть любое другое слово, выражающее профессию. При этом нужно учитывать наличие между этим словом и ФИО факультативных элементов, например, названия организации. Таким образом начинается выделение подразумеваемых объектов, т.е. тех, у которых нет характеристических слов, определяющих их наличие. Например, не распознаны компоненты ФИО.

В текстах ЕЯ многие связи подразумеваются и привязаны к типу выявленных объектов. Например, если выявлен адрес, то, скорее всего, он относится к какому-либо определенному лицу (или организации), которое нужно искать. При результативном поиске формируется новая связь. На этом основана методика формирования новых связей. Она заключается в следующем. В процессе анализа текста строятся «временные» фрагменты, представляющие связи выявленных объектов с пока что неизвестными объектами, которые специальным образом отмечаются. В дальнейшем

осуществляется их поиск. Если соответствующий объект не найден, то «временный» фрагмент удаляется из СП-документа. Если найден, то фрагмент остается – вводится в структуру СП-документа.

Аналогичная методика используется при формировании новых признаков. Формируется признак с пока что неизвестным объектом, который в дальнейшем уточняется.

При формировании объектов некоторые компоненты могут быть сразу не найдены, например, год рождения, который в СП-документа представляется как компонента ФИО. Тогда в соответствующих фрагментах специальными константами отмечаются незаполненные аргументные места, которые в дальнейшем уточняются. Для более детального описания методик и средств их реализации рассмотрим правила и этапы построения СП-документов в процессе синтактико-семантического анализа.

Отметим, что при глубинном анализе текстов (выделении действий и их участников) определенные трудности вызывает наличие в анализируемых глагольных формах словосочетаний, представляющих причину действий («на почве неприязненных отношений», «в ссоре», «из хулиганских побуждений», …), сопутствующие действия («при личном досмотре», «при поставке оружия», «во время кражи», …) и др. Многие из таких словосочетаний в сводках происшествий встречаются регулярно и поэтому задаются в виде перечней – в соответствующем предметном словаре.

4 Правила синтактико-семантического анализа

Синтактико-семантический анализ необходим для выделения связанных групп слов, а также информационных объектов («именованных сущностей»): адресов, номеров машин, организаций и др. Последние, как правило, это наборы слов, которые могут быть грамматически никак не согласованы. Их выделение осуществляется по чисто формальным принципам на основе правил, составляющих ЛЗ. Например, адрес может рассматриваться как набор буквосочетаний «г.», «ул.», «д.», …, слов с большой буквы и чисел. Каждый такой набор может иметь свои границы и недопустимые компоненты. Например, в адресах не может быть местоимений, глаголов и т.д. Выделение таких наборов слов, составляющих описания объектов, основано на использовании правил синтактико-семантического анализа (в дальнейшем просто – правил) следующего вида:

<ПравилоN>:CONTEXT(<слово1>,<слово2>, ...) --> <результирующий фрагмент>,

где <ПравилоN> – имя правила, необходимое для его вызова, а <слово1>, <слово2>, ... – это может быть отдельное слово, признак, а также И-ИЛИ граф, составленный из слов и признаков. Для этих правил указывается, с какой позиции начинать применение, а также допустимый или недопустимый контекст. Обычно применение начинается с позиции, на которой находятся характеристические слова. Например, выделение лиц начинается с поиска распознанных компонент ФИО. Выделение адресов – с поиска слов: ул., дом, кв. и т.д.

Правила выделяют из текста группы слов (по их признакам), описывающих какой-либо объект, и заменяют их на одно (абстрактное) слово, с которым связывается соответствующий фрагмент семантической сети и которому присваиваются определенные признаки, в том числе признак, указывающий на тип объекта.

Синтактико-семантический анализ предложений (с выделением словосочетаний и анализом языковых конструкций) осуществляется на основе правил, которые применяются в определенной последовательности. Вначале выделяются простейшие объекты, затем согласованные группы слов, затем более сложные объекты и их признаки, и, наконец, глагольные формы, п. 4. По мере применения таких правил строится семантическая сеть – содержательный портрет документа. Например, рассмотрим правило с именем GG~1:

MUSTBE(GG~1,1) STR_OR(ADJ,PRON/2+) CONTEXT(2-,NOUN/GG~1) P_P(GG~1,3+) WORD_C(1,2/3-) NOTBE(GG~1,2,LETT).

Правило GG~1 осуществляет преобразования:

GG~1:ПРИЛАГАТЕЛЬНОЕ + СУЩЕСТВИТЕЛЬНОЕ --> <комбинация слов>

МЕСТОИМЕНИЕ + СУЩЕСТВИТЕЛЬНОЕ --> <комбинация слов>.

Фрагмент MUSTBE указывает, что применять правило GG~1 нужно с 1-ой позиции, т.е. искать слова с признаками ПРИЛАГАТЕЛЬНОЕ (ADJ) и МЕСТОИМЕНИЕ (PRON), так как их меньше, чем СУЩЕСТВИТЕЛЬНЫХ (NOUN). Символ 2+ – это код фрагмента типа «ИЛИ» (STR_OR), а фрагмент CONTEXT(2-,NOUN/GG~1) задает позиции правила GG~1, где на первой позиции стоит указанный код (его повторное применение обозначается 2-), а на второй – признак NOUN. Аналогичным образом используются символы 3+ и 3-.

Фрагмент P_P отделяет левую часть от правой (- ->), а WORD_C – указывает, что слова на 1-й и 2-ой позициях должны быть склеены в комбинацию слов, которая в дальнейшем будет рассматриваться как одно слово с морфологическими признаками 2-го слова. Фрагмент NOTBE указывает, что на 2-ой позиции не могут быть отдельные буквы (признак LETT). К данному правилу добавляется фрагмент, требующий согласованности слов (по падежам, числам), а также фрагменты, задающие с признаков и контекстные ограничения.

Это пример наиболее простого правила. Более сложные правила, построенные по аналогичным принципам, осуществляют выделение сложных объектов и действий. Помимо этого, в ЛЗ имеются специальные правила, которые осуществляют идентификацию объектов, например, с местоимениями или краткими описаниями (по имени восстанавливается фамилия, если они где-нибудь упоминались вместе). И многое другое, что необходимо при автоматическом построении СП-документа, отражающем семантически значимые компоненты ЕЯ-текста.

Отметим, что каждое правило (как и все лингвистические знания) записывается на языке РСС и является частью ЛЗ. Над правилами работают продукции языка ДЕКЛ (программа), которые применяют эти правила и играют роль пустой лингвистической оболочки, поддерживающей язык записи лингвистических знаний - РСС. Как показывает опыт, такую оболочку можно настраивать на различные языки, т.е. строить различные лингвистические процессоры, в том числе, англоязычные [6], также (http://www.ipiranlogos.com/english/topics/topic3-e.htm).

5 Порядок применения правил

Правила синтактико-семантического анализа применяются в строго определенной последовательности – каждое на своем уровне. Например, при обработке сводок происшествий вначале выделяются информационные объекты – отделения милиции (ОВД_), сотрудники милиции (МИЛ_) и др. Они могут содержать фамилии, имена, которые следует отличать от ФИО лиц – фигурантов (последние представляются фрагментами FIO). Далее выделяются статьи УК и т.д. Это необходимо, чтобы облегчить последующий анализ. Иначе слова, составляющие эти объекты, могут захватываться другими правилами и создавать шумы.

Далее начинается выделение лиц – фигурантов. Для этого вводится множество правил. Одни правила начинают свое применение с поиска распознанных имен или фамилий (MUSTBE), другие – с поиска года рождения, третьи – с инициалов. В результате минимизируются потери в случаях, когда блок морфологического анализа не дает необходимых признаков для каких-либо слов (что это имена или фамилии и т.д.).

Затем анализируются словосочетания, выделяются объекты, и наконец, анализируются глагольные формы. По мере применения таких правил строится СП-документа. Последовательность правил задается с помощью специальных фрагментов. Ниже приведен пример представления уровней, определяющих порядок применения правил.

```
{== Уровни ==}
 LEVEL(LEVEL1,LEVEL2,LEVEL3,LEVEL4,...)
 LEVEL1(CATALOG)   {= Объединение словосочетаний из каталогов =}
 . . .
 LEVEL2(MIL~1,ST~1)         {= Выявление отд.милиции, ст. УК =}
 LEVEL3(DD~1,DD~2,…)     {= Выявление времени, дат, в том числе, г.рожд. =}
 LEVEL4(FF~1,FF~2, …)       {= Выявление лиц с распознанными ФИО =}
 LEVEL4(FA~1)             {= Выявление нераспознанных лиц =}
 LEVEL4(ID_4)              LEVEL11(ID_2A,ID_2,ID_21)      {=      идентификация
местоимений =}
        {= Поиск года рождения для выявленных лиц =}
 LEVEL4(PROP~1,PROP~2,ID_33)  {= Выявление свойств и поиск лиц =}
 . . .
 LEVEL5(AA~1,AA~2)         {= Выявление однородных членов =}
 LEVEL6(GG~1,GG~2,...)      {= Выявление словосочетаний =}
. . .
 LEVEL10(ID_1)                  {= идентификация связок «тот, который» =}
 LEVEL11(ID_2A,ID_2,ID_21) {= идентификация местоимений =}
 . . .
 LEVEL13(CL~1,CL~2A, …)    {= Выявление адресов =}
 . . .
 LEVEL15(VV~1, …) {= Выявление действий (анализ глагольных и др. форм) =}
 . . .
```

В фигурных скобках даны комментарии. Первый фрагмент LEVEL(…) задает уровни, а последующие – правила каждого уровня.

Правила начинают применяться к семантической сети (ПС-текста), которая имеет вид линейной структуры и в которой последовательность слов задается с помощью фрагментов LR. С ними связываются распознанные признаки слов: лексические, морфологические, семантические. Предложения разделяются фрагментами SENT. Все это представляется на РСС.

Правила анализируют линейную структуру, находят соответствующие группы слов, из которых формируются объекты. При этом объекты как бы замещают эти слова. Линейная структура сохраняется, но видоизменяется. В конце остается линейная структура (на РСС), компонентами которой являются объекты и слова, не вошедшие в объекты (напомним, что события и действия – это тоже объекты). На этой основе формируется СП-документа [5], [6].

В ЛП имеются правила, которые обеспечивают полный разбор предложений. При этом параллельно обеспечивается выделение необходимого набора информационных объектов, в том числе таких, в которых слова никак не согласованы между собой, например, адресов, машин с указанием их номеров и т.д. [3], [4]. На рис. 2 представлен процесс применения правил синтактико-семантического анализа.

На рис. 2 после звездочек (* *) показан текст, на основе которого построен информационный объект. Правило CL~2A выделяет адрес, а правило ID~31 связывает этот адрес отношением ПРОЖ (проживать) с фигурантом – Митрофановым Виктором Михайловичем. Правило VV~1 осуществляет анализ глагольных форм, выделяет действия и тех, кто в них участвует.

```
Логико-аналитическая система
Файл   Сервис   Справка
Выход   Запрос-ЕЯ   Задачи   Каталоги   Настройки   База-знаний
Введено документов - 26
F1  F4  F5  F7  F8  F9  ALT-F9
ДАТА_(#01.02.98,98,02,~01,"10-00"/31+) ->     << TT~4
 ДАТА :   #01.02.98 98 02 ~01 10-00    << TT~4
** 01.02.98в 10 - 00
 +++ Уровень +++ LEVEL4
 WORD_C:   НЕЦЕНЗУРНЫЙ ГРУБОСТЬ    << OO~1
** нецензурной бранью
 WORD_C:   УКУС НОГА    << GG~2
** укус ноги
 +++ Уровень +++ LEVEL5
 АДР_:     БОРОВСКИЙ ШОССЕ 38 - 211    << CL~2A
** прож. : Боровское шоссе 38 - 211
 ПРОЖ.:  МИТРОФАНОВ ВИКТОР МИХАЙЛОВИЧ 1955   БОРОВСКИЙ ШОССЕ 38 - 211     << ID_31
 АДР_:  ДОМ 3   УЛ.   ФЕДОСЬИНО     << CL~1
** д. 3 по ул. Федосьино
 ИМЕТЬ:  МИТРОФАНОВ ВИКТОР МИХАЙЛОВИЧ 1955 ДОМ 3   УЛ   ФЕДОСЬИНО   MT_ID_1    << ID_3
 ОН - это  ФИО:  МИТРОФАНОВ ВИКТОР МИХАЙЛОВИЧ 1955   << ID_2A(PRON_1,32+,33+,30,LEFT)
 +++ Уровень +++ LEVEL6
Когда.  ДАТА_.  #01.02.98 98 02 ~01 16-30
 ОБРАТИТЬСЯ:    В   ОВД          МИТРОФАНОВ ВИКТОР МИХАЙЛОВИЧ 1955          << VV~1
**  в ОВД обратился гр-н Митрофанов Виктор Михайлович , 1955 год р.
 СКАНДАЛ:        НЕСКОЛЬКО   МИТРОФАНОВ ВИКТОР МИХАЙЛОВИЧ 1955           << VV~
1
** неизвестные находясь в пьяном виде    учинили скандал
 ВЫРАЖАТЬСЯ:           НЕЦЕНЗУРНЫЙ ГРУБОСТЬ           << VV~1
** выражались нецензурной бранью
 НАТРАВИТЬ:           СОБАКА              << VV~1
** натравили собаку
-> Запрос на ЕЯ
```

Рисунок 2 – Процесс применения правил

6 Принцип «ожидания» при выявлении объектов

При наличии в тексте объектов без характеристических слов возникают трудности их выделения. Например, если в тексте встречаются лица с иностранными ФИО. У английских фамилий («Буш», «Райс», «Браун», …) нет характерных суффиксов, как в русском языке. Более того, в качестве фамилий может быть любое слово, называющее или определяющее какой-либо предмет внешнего мира. При анализе англоязычных текстов такие фамилии вносят элементы неопределенности – омонимии. В азиатских языках компоненты ФИО – это просто слова с большой буквы («Ден Сяо Пин», «Лю Шао Ци», …). Задать перечислением все данные имена или фамилии (в предметных словарях) не представляется возможным. В подобных ФИО отсутствуют характеристические слова. Требуются другие методики выделения. Аналогично, адреса могут иметь вид – «Никольская 12-55». Сказанное относится и к другим объектам.

Для выделения, как уже говорилось, используется принцип «ожидания» – после одних объектов (или понятий) ожидается наличие других. Реализация соответствующей методики осуществляется с помощью операторов вида:

GO_(<Правило1>,<Правило2>,N),

где Правило 1 – правило, которое было вызвано. И если оно применилось, то оно вызывает Правило 2, применение которого начинается с позиции N.

Рассмотрим пример использования данного оператора при выявлении ФИО. Это осуществляется с помощью двух правил – FA~1 и FF~1:

```
MUSTBE(FA~1,1) STR_OR(WORK_K,NAT_K/2+) CONTEXT(2-/FA~1)
    P_P(FA~1,“ ”) GO_(FA~1,FF~1,1).

MUSTBE(FF~1,1) STR_OR(NAME0/3+) CONTEXT(3-,3-,3-/FF~1)
    P_P(FF~1,4+) FIO(1,2,3,“ ”/4-)  MAYBE(FF~1,3)
    STR_OR(VERB,ENG/5+) NOTBE(FF~1,ALL,5-).
```

Правило FA~1 находит в тексте слова с признаками WORK_K (профессии) и NAT_K (национальность). Такие признаки присваиваются словам блоком морфологического анализа на основе предметных словарей, где даны списки профессий, национальностей и др. [4]. И если слово с таким признаком найдено, то вызывается правило FF~1, которое проверяет, чтоб за найденным словом стояли 3 слова с большой буквы (с признаком NAME0). При этом такие слова не могут быть (NOTBE) глаголами (их признак VERB) или англоязычными (их признак – ENG), что задается с помощью двух последних фрагментов. Фрагмент MAYBE(FF~1,3) указывает, что третья позиция является факультативной, т.е. третьего слова с большой буквы (ББ) может не быть. И всего одно правило будет применимым. В случае применимости формируется фрагмент FIO(…). У него в качестве первых трех аргументов будут первые три слова, которые удовлетворяют условиям, заданным в фрагменте CONTEXT. Эти три слова заменяются на одно, с которым связывается сформированный фрагмент и к которому добавляется признак FIO.

Эти два правила осуществляют преобразования:

ПРОФЕССИЯ + 2 или 3 СЛОВА С ББ --> <выделенное лицо>,

НАЦИОНАЛЬНОСТЬ + 2 или 3 СЛОВА С ББ --> <выделенное лицо>.

Например, словосочетание «председатель Ху Цзинь Тао» будет преобразовано в фрагмент FIO (ХУ, ЦЗИНЬ, ТАО, ” ”). При этом слово «председатель» останется и будет использовано при последующем анализе. Словосочетание «премьер Хапер Стивен» будет преобразовано в фрагмент FIO (ХАПЕР, СТИВЕН, ” ”, ” ”). Для выделения ФИО из словосочетаний типа «премьер Канады Хапер Стивен» в фрагмент CONTEXT первого правила необходимо вставить факультативную позицию для слов с признаком «государство». Путем модификации правил можно охватить множество случаев, не увеличивая количество правил.

Другой способ выделения ФИО – через глаголы, субъектами которых могут быть только лица. Например, «…Хапер Стивен подписал…», где глагол «подписать» помогает выделению лица. Такие глаголы даются перечнем («предложить», «подписать», «согласиться», …), а выделение лиц реализуется с помощью того же оператора GO_.

Отметим, что правила выделения объектов и правила идентификации представлены в лингвистических знаниях в виде наборов элементарных фрагментов РСС, которые легко менять, настраивая лингвистический процессор (ЛП) на ту или иную предметную область. Сама программа (на языке ДЕКЛ) остается неизменной. Этот фактор дает большие преимущества при отладке и настройке ЛП, так как учесть даже малую часть того, что может встретиться в ЕЯ, не представляется возможным. ДЕМО-версию процессора ЛП можно найти на сайте (http://ipiranlogos.com/ru/Demo-1/).

Заключение

В данной статье рассмотрены семантические методики по извлечению структур знаний из текстов естественного языка. Предлагаемые методики реализованы в рамках единого инструментального комплекса: языка расширенных семантических сетей (РСС) для представления знаний и средств их обработки – языка ДЕКЛ. Этот комплекс ориентирован на организацию баз знаний и на их использование для решения интеллектуальных задач, в том числе, связанных с извлечением структур знаний, их анализом для дополнения и корректировки структур, логическим выводом, принятием экспертных решений. Предметные и лингвистические знания представляются на единой основе (в виде фрагментов РСС), что позволяет свести казалось бы разнородные задачи к преобразованию структур знаний. Это дает определенные преимущества: упрощает создание соответствующих программ (на языке ДЕКЛ), обеспечивающих анализ высокой степени глубины и сложности.

Технологический комплекс обладает следующими особенностями:

1 Из текстов извлекаются не отдельные объекты (именованные сущности), а структуры знаний, представляющие связи объектов и их участие в действиях и событиях.

2 Для извлечения структур знаний разработан уникальный семантико-ориентированный лингвистический процессор (ЛП), осуществляющий глубинный анализ текстов ЕЯ и выявляющий десятки типов объектов вместе с их структурами.

3. Процессор ЛП управляется лингвистическими знаниями, представляющими собой декларативные структуры (на РСС) и обеспечивающие быструю настройку ЛП на предметную область и язык.

4. Основой лингвистических знаний являются правила, обладающие высокой степенью избирательности при выявлении объектов («сущностей»), средствами устранения коллизий при их применении. Это позволяет минимизировать шумы и потери – добиваться высокой степени полноты и точности.

5. Структуры знаний (на РСС) и средства их обработки (язык ДЕКЛ) разрабатывались как единый инструментарий, ориентированный на задачи лингвистического анализа, семантического поиска, логико-аналитической обработки и экспертных решений. Использование этого инструментария значительно облегчает разработку лингвистических процессоров и прикладных интеллектуальных систем.

Литература

1. Byrd R. Identifying and Extracting Relations in Text / R. Byrd and Y. Ravin // 4th International Conference on Applications of Natural Language to Information Systems (NLDB). – Klagenfurt, Austria, 1999.
2. Open Information Extraction from the Web / M. Banko, M. Cafarella, S. Soderland [and al.] // Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07), 2007. – P. 2670-2676.
3. Кузнецов И.П. Семантико-ориентированные системы на основе баз знаний : [монография] / И.П. Кузнецов, А.Г. Мацкевич. – М. : МТУСИ, 2007 г. – 173 с.
4. Особенности лексико-морфологического анализа в задачах извлечения структур знаний из текстов естественного языка / И.П. Кузнецов, Н.В. Сомин, Е.Б. Козеренко [и др.] // Искусственный интеллект, НАН Украины, 2011. – Т. 3. – С. 105-116.
5. Кузнецов И.П. Принципы организации объектно-ориентированных систем обработки неформализованной информации / И.П. Кузнецов, Е.Б. Козеренко, А.Г. Мацкевич // Искусственный интеллект, НАН Украины, ИПИИ. – 2010. – Вып. 3. – С. 227-237.
6. Kozerenko E.B. The system for extracting semantic information from natural language texts / Eliena B. Kozerenko, Igor P. Kuznetsov // Proceeding of International Conference on Machine Learning. MLMTA-03, Las Vegas US, 23-26 June 2003. – P. 75-80.
7. Сайт «Интеллектуальные системы обработки знаний» [Электронный ресурс]. – Режим доступа : http://IpiranLogos.com
8. Kuznetsov I.P. Linguistic Processor «Semantix» for Knowledge extraction from natural texts in Russia and English / Igor P. Kuznetsov, Elena B. Kozerenko // Proceeding of International Conference on Machine Learning, ISAT-2008. 14 – 18 July, 2008. – Las Vegas, USA CSREA Press, 2008. – P. 835-841.
9. Kuznetsov I.P. Intelligent extraction of knowledge structures from natural language texts / I.P. Kuznetsov, E.B. Kozerenko, A.G. Matskevich // Proceedings-2011 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology. – Workshops, WI-IAT 2011. – P. 269-272.
10. Кузнецов И.П. Семантические методы извлечения имплицитной информации / Кузнецов И.П. : сб. Системы и средства информатики. – Вып.21, № 2. – М. : Наука, 2011. – С. 116-138.

Literatura

1. Byrd R. Identifying and Extracting Relations in Text / R. Byrd and Y. Ravin // 4th International Conference on Applications of Natural Language to Information Systems (NLDB). – Klagenfurt, Austria, 1999.
2. Open Information Extraction from the Web / M. Banko, M. Cafarella, S. Soderland [and al.] // Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07), 2007. – P. 2670-2676.

3. Кузнецов И.П. Семантико-ориентированные системы на основе баз знаний : [монография] / И.П. Кузнецов, А.Г. Мацкевич. – М. : МТУСИ, 2007 г. – 173 с.
4. Особенности лексико-морфологического анализа в задачах извлечения структур знаний из текстов естественного языка / И.П. Кузнецов, Н.В. Сомин, Е.Б. Козеренко [и др.] // Искусственный интеллект, НАН Украины, 2011. – Т. 3. – С. 105-116.
5. Кузнецов И.П. Принципы организации объектно-ориентированных систем обработки неформализованной информации / И.П. Кузнецов, Е.Б. Козеренко, А.Г. Мацкевич // Искусственный интеллект, НАН Украины, ИПИИ. – 2010. – Вып. 3. – С. 227-237.
6. Kozerenko E.B. The system for extracting semantic information from natural language texts / Eliena B. Kozerenko, Igor P. Kuznetsov // Proceeding of International Conference on Machine Learning. MLMTA-03, Las Vegas US, 23-26 June 2003. – P. 75-80.
7. Сайт «Интеллектуальные системы обработки знаний» [Электронный ресурс]. – Режим доступа : http://IpiranLogos.com
8. Kuznetsov I.P. Linguistic Processor «Semantix» for Knowledge extraction from natural texts in Russia and English / Igor P. Kuznetsov, Elena B. Kozerenko // Proceeding of International Conference on Machine Learning, ISAT-2008. 14 – 18 July, 2008. – Las Vegas, USA CSREA Press, 2008. – P. 835-841.
9. Kuznetsov I.P. Intelligent extraction of knowledge structures from natural language texts / I.P. Kuznetsov, E.B. Kozerenko, A.G. Matskevich // Proceedings-2011 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology. – Workshops, WI-IAT 2011. – P. 269-272.
10. Кузнецов И.П. Семантические методы извлечения имплицитной информации / Кузнецов И.П. : сб. Системы и средства информатики. – Вып.21, № 2. – М. : Наука, 2011. – С. 116-138.

RESUME

Igor P. Kuznetsov, Mikhail M. Sharnin, Andrey G. Matskevich

Technology of Knowledge Extraction on the base of Extended Semantic Networks

The paper devoted the task of extracting from the texts of natural language structures of knowledge: information objects («Named Entity»), their properties, relationships, and participation in the actions and events. For this purpose, the language used for knowledge representation (extended semantic networks – RCC) and tools for processing (language structure conversion – DCL) are considered. On this base the new technologies are proposed. Distinctive features of our technology:

1 Extraction from the texts of knowledge structures that represent the links of named entities and their participation in actions and events.

2 For the knowledge extraction the unique semantic-oriented language processor (LP) are designed. Processor LP provides the deep analysis of NL-texts and revealing set of objects together with their structures.

3 Processor LP is controlled by the linguistic knowledge, which are declarative structures (on extended semantic networks - ESN) and which provides the quick tuning of LP on subject area and language – Russian and English.

4 Linguistic knowledge consists of the rules, which provide the high degree of selectivity in the entities extraction and elimination of collisions during their application. Rules provide the minimization of noise and losses, that is the high degree of completeness and accuracy.

5 The knowledge structures and means of their processing (intellectual language DEKL) were designed as the united tools, oriented to the tasks of linguistic analysis, semantic search, logical-analytical processing and the expert solutions.

The using this tools considerably facilitates the designing applied intellectual systems.

Статья поступила в редакцию 01.06.2012.